Claims:

1. A method of estimating selectivity of a given string predicate in a database query, comprising:

a) estimating selectivities of string predicate substrings of various substring lengths;

b) selecting a candidate substring for each substring length based on estimated selectivities of the substrings;

c) combining the estimated selectivities of the candidate substrings; and

d) returning the combined estimated selectivities of the candidate substrings as the estimated selectivity of the given string predicate.

2. The method of claim 1 further comprising storing selectivity information for the database and using stored selectivity information to estimate the selectivities of the substrings of various lengths.

3. The method of claim 1 wherein a substring with a lowest estimated selectivity is selected as the candidate substring at each length.

4. The method of claim 1 further comprising calculating exact selectivities of substrings of a given maximum length and using the exact selectivities to estimate the selectivities of the substrings of various substring lengths.

5. The method of claim 4 wherein a range of the various substring lengths whose selectivities are estimated is between the given maximum length of the substrings whose selectivities are calculated exactly and the length of the given string predicate.

6. The method of claim 4 wherein the candidate substring for the length equal to the given maximum length of the substrings whose selectivities are calculated exactly is selected based on the exact selectivity of the substring.

7. The method of claim 1 wherein a q-gram table is constructed for substrings of a given maximum length and is accessed to estimate selectivities of substrings of various substring lengths.

8. The method of claim 4 wherein a markov estimator uses the exact selectivities to estimate the selectivities of the substrings of various substring lengths.

9. The method of claim 1 wherein characteristics of string values in a relation of the database are used to combine the estimated selectivities of the candidate substrings.

10. The method of claim 1 wherein characteristics of a workload of queries are used to combine the estimated selectivities of the candidate substrings.

11. The method of claim 1 wherein a model for combining the estimated selectivities of candidate substrings is learned from query workloads.

12. The method of claim 1 wherein said model is applied to the candidate substrings at run time to estimate the string predicate selectivity.

13. The method of claim 1 wherein the given string predicate is a unit predicate.

14. The method of claim 1 wherein the given string predicate includes a wildcard character.

15. The method of claim 1 wherein the given string predicate is a range predicates.

16. The method of claim 1 wherein weights are assigned to each length of candidate substring to combine the selectivities of the candidate substrings.

17. The method of claim 16 wherein a function for assigning said weights is learned from data sets of the database.

18. The method of claim 16 wherein a function for assigning said weights is learned from an expected query workload.

19. The method of claim 16 further comprising calculating actual selectivities of substrings of queries from an expected workload and determining estimated selectivities

of the substrings of a queries from the expected workload to learn a function for assigning said weights.

20. The method of claim 16 further comprising calculating for a string predicate of a query from an expected workload an actual selectivity of a candidate substring having the given length, determining for the string predicate of the query from the expected workload an estimated selectivity of the candidate substring having the given length, and assigning a weight to candidate substrings of a given length by based on a relationship between the calculated actual selectivity and the determined estimated selectivity.

21. The method of claim 1 wherein selectivities of the candidate substrings are combined using regression trees.

22. The method of claim 20 wherein said regression trees are learned from data sets of the database.

23. The method of claim 20 wherein said regression trees are learned from an expected query workload.

24. A computer readable medium having computer executable instructions stored thereon for performing a method of estimating selectivity of a given string predicate in a database query, the method comprising:

a) estimating selectivities of substrings of various substring lengths;

b) selecting a candidate substring for each substring length based on estimated selectivities of the substrings;

c) combining the estimated selectivities of the candidate substrings; and

d) returning the combined estimated selectivities of the candidate substrings as the estimated selectivity of the given string predicate.

25. The computer readable medium of claim 24 wherein the method further comprises storing selectivity information for the database and using stored selectivity information to estimate the selectivities of the substrings of various lengths.

26. The computer readable medium of claim 24 wherein a substring with a lowest estimated selectivity is selected as the candidate substring at each length.

27. The computer readable medium of claim 24 wherein the method further comprises calculating exact selectivities of substrings of a given maximum length and using the exact selectivities to estimate the selectivities of the substrings of various substring lengths.

28. The computer readable medium of claim 27 wherein a range of the various substring lengths whose selectivities are estimated is between the given maximum length of the substrings whose selectivities are calculated exactly and the length of the given string predicate.

29. The computer readable medium of claim 27 wherein the candidate substring for the length equal to the given maximum length of the substrings whose selectivities are calculated exactly is selected based on the exact selectivity of the substring.

30. The computer readable medium of claim 24 wherein a q-gram table is constructed for substrings of a given maximum length and is accessed to estimate selectivities of substrings of various substring lengths.

31. The computer readable medium of claim 27 wherein a markov estimator uses the exact selectivities to estimate the selectivities of the substrings of various substring lengths.

32. The computer readable medium of claim 24 wherein characteristics of string values in a relation of the database are used to combine the estimated selectivities of the candidate substrings.

33. The computer readable medium of claim 24 wherein characteristics of a workload of queries are used to combine the estimated selectivities of the candidate substrings.

34. The computer readable medium of claim 24 wherein a model for combining the estimated selectivities of candidate substrings is learned from query workloads.

35. The computer readable medium of claim 24 wherein said model is applied to the candidate substrings at run time to estimate the string predicate selectivity.

36. The computer readable medium of claim 24 wherein the given string predicate is a unit predicate.

37. The computer readable medium of claim 24 wherein the given string predicate includes a wildcard character.

38. The computer readable medium of claim 24 wherein the given string predicate is a range predicates.

39. The computer readable medium of claim 24 wherein weights are assigned to each length of candidate substring to combine the selectivities of the candidate substrings.

40. The computer readable medium of claim 39 wherein a function for assigning said weights is learned from data sets of the database.

41. The computer readable medium of claim 39 wherein a function for assigning said weights is learned from an expected query workload.

42. The computer readable medium of claim 39 wherein the method further comprises calculating actual selectivities of substrings of queries from an expected workload and determining estimated selectivities of the substrings of a queries from the expected workload to learn a function for assigning said weights.

43. The computer readable medium of claim 39 wherein the method further comprises calculating for a string predicate of a query from an expected workload an actual selectivity of a candidate substring having the given length, determining for the string predicate of the query from the expected workload an estimated selectivity of the candidate substring having the given length, and assigning a weight to candidate substrings of a given length by based on a relationship between the calculated actual selectivity and the determined estimated selectivity.

44. The computer readable medium of claim 24 wherein selectivities of the candidate substrings are combined using regression trees.

45. The computer readable medium of claim 44 wherein said regression trees are learned from data sets of the database.

46. The computer readable medium of claim 44 wherein said regression trees are learned from an expected query workload.